Object Detection and Identification for Blind People in Video Scene

Hanen Jabnoun[1], Faouzi Benzarti[1], Hamid Amiri[1]
[1] LR-11-ES17 Signal, Images et Technologies de l'Information (LR-SITI-ENIT)
Université de Tunis El Manar, Ecole Nationale d'Ingénieur de Tunis
1002, Tunis Le Belvédère, Tunisie
hanen.jabnoun@enit.rnu.tn benzartif@yahoo.fr hamid.amiri@enit.rnu.tn

Abstract— Vision is one of the very essential human senses and it plays the most important role in human perception about surrounding environment. Hence, over thousands of papers have been published on these subjects that propose a variety of computer vision products and services by developing new electronic aids for the blind. This paper aims to introduce a proposed system that restores a central function of the visual system which is the identification of surrounding objects. This method is based on the local features extraction concept. The simulation results using SFIT algorithm and keypoints matching showed good accuracy for detecting objects. Thus, our contribution is to present the idea of a visual substitution system based on features extractions and matching to recognize and locate objects in images.

Keywords—Video processing; pattern recognition; SIFT; keypoints matching; visual substituion system.

I. INTRODUCTION

According to the World Health Organization, there are approximately 285 million people who are visual impairments, 39 million of them are blind and 246 million have a decrease of Visual acuity. Almost 90% who are visually impaired are living in low-income countries. In this context, Tunisia has identified 30,000 people with visual impairments; including 13.3% of them are blind.

These Visual impairment present severe consequences on certain capabilities related to visual function:

- The daily living activities (that require a vision at a medium distance)
- Communication, reading, writing (which requires a vision closely and average distance)
- Evaluation of space and the displacement (which require a vision far)
- The pursuit of an activity requiring prolonged maintenance of visual attention.

In the computer vision community, developing visual aids for handicapped persons is one of the most active research projects. Mobility aids are intended to describe the environment close to the person with an appreciation of the surrounding objects. These aids are essential for fine navigation in an environment described in a coordinate system relative to the user. In this paper, we present an overview of vision substitution modalities [1-12] and their functionalities. Then, we introduce our proposed system and the experiments tests.

II. OVERVIEW

Related works show that visual substitution devices accept input from the user's surroundings, decipher it to extract information about entities in the user's environment, and then transmit that information to the subject via auditory or tactile means or some combination of these two.

Among the various technologies used for blind people, the majority is aids of mobility and obstacle detection [5, 8].

They are based on rules for converting images into data sensory substitution tactile or auditory stimuli. These systems are efficient for mobility and localization of objects which is sometimes with a lower precision.

However, one of the greatest difficulties of blind people is the identification of their environment and its [6]. Indeed, they can only be used to recognize simple patterns and cannot be used as tools of substitution in natural environments. Also, they don't identify objects (e.g. whether it is a table or chair) and they have in some cases a late detection of small objects. In addition, some of them seek additional auditory, others require a sufficiently long period for learning and testing.

Among the problems in object identification, we note the redundancy of objects under different conditions: the change of viewpoint, the change of illumination and the change of size. We have the concept of intra-class variability (e.g. there are many types of chairs) and the inter-class similarity (e.g. television and computer).

For this reason, we are interested in the evaluation of an algorithm for fast and robust computer vision application to recognize and locate objects in a video scene.

Thus, it is important to design a system based on the recognition and detection of objects to meet the major challenges of the blind in three main categories of needs: displacement, orientation and object identification.

III. OBJECT DETECTION BASED ON FEATURES EXTRACTION

Object recognition is a classical problem in computer vision: the task of determining if the image data contains a specific object and it is noted that general object recognition approaches exploit features extraction. Features that have received the most attention in the recent years are the local features. The main idea is to focus on the areas containing the most discriminative information.

A. Features extraction

The main purpose of using features instead of raw pixel values as the input to a learning algorithm is to reduce/increase the in-class/out-of class variability compared to the raw input data, and thus making classification easier [9]. Visual substitution systems generally exploit a single camera to capture image data. Recognition is then performed based on various features extracted from that data. In addition, features extraction is the process by which certain features of interest within an image are detected and represented for further processing. It is a critical step in most computer vision and image processing solutions because it marks the transition from pictorial to non-pictorial (alphanumerical, usually quantitative) data representation [13]. Types of features that can be extracted from image depend on the type of image, the level of granularity desired, and the context of the application. Once the features have been extracted, theirs representation depends on the technique used.

The features extraction process should be precise, so that the same features are extracted on two images showing the same object [12]. The major algorithm used in recent research for features extraction are the Scale Invariant Features Transform (SIFT) [13] and the Speed Up Robust Features (SURF) [15] algorithms.

B. Features descriptors

Extracted features represent the interesting points found in the image to compare them with other interesting points. Descriptors are used to describe these features. They are generally based around points of interest of the image and often associated with a detector of keypoints [10]. The descriptors can be global, local or semi-local:

- Global image descriptor: features overall image are usually based on color indices and the most famous global color descriptor is the color histogram.
- Local image descriptors: Local features are ones that have received the most attention in recent years. The main idea is to focus on the areas containing the most discriminated information.
- Semi-local image descriptor: most shape descriptors fall into this category. This descriptor is based on extracting accurate contours of shapes in the image or in the region of interest. In this case, image segmentation is generally useful as a preprocessing step.

C. SIFT keypoints extraction

Several methods for features localization and description have been proposed in the literature. In this paper, we aim to use local features extraction algorithm SIFT.

The SIFT (Scale Invariant Feature Transform) local descriptor proposed by Lowe [13] has been one of the most widely used descriptors.

It operates in four major stages to detect and describe local features, or keypoints, in an image:

1. Detect the extrema in scale space
2. Sub-unit localization and filtering of keypoints
3. Assign the orientations to keypoints
4. Compute the keypoint descriptors

In fact, SIFT was shown to perform better than all other local descriptors. SIFT descriptor is based on the idea of using the local gradient patch around the keypoint to build a representation for the point.

Different levels of Gaussian filtered image are subtracted to build the difference of Gaussian image for further calculation. Differences of Gaussian (DoG) for each octave find extrema from DoG. Each octave is computed by convolution between down sampled images with different Gaussian kernels

The Difference of Gaussian is obtained from the difference of two adjacent scales that are separated by a factor of k [14].

$$D(X,\sigma) = (G(X,k\sigma) - G(X,\sigma)) * I(X) \quad (1)$$

Where:

- I: the input image.
- X: the point X(x ,y).
- σ: the scale
- $G(X,\sigma)$: The variable-scale Gaussian.

The DoG interest regions are defined as locations that are simultaneously extrema in the image plane and along the scale coordinate of the D(x, σ) function (1). Such points are found by comparing the D(x, σ) value of each point with its 8-neighborhood on the same scale level, and with the 9 closest neighbors on each of the two adjacent levels.

For each feature point in the image, a square patch around it is appropriately scaled and rotated. Then, the 8 bins orientation histograms over a 4x4 region are generated using the gradient magnitudes and orientations in the patch. Finally, the outputs of 16 orientation histograms which results in a 128 element feature descriptor are computed to get the final descriptor.

IV. PROPOSED APPROACH FOR VISUAL SUBSTITUTION SYSTEM

For objects recognition in videos, a lot of work has been done using so-called feature extraction and detection along the video or by extraction of visually salient regions which are supposed to contain object of interest.

Feature matching using invariant features has obtained significant importance due to its application in various recognition problems. Such techniques have enabled us to match images irrespective of various geometric and photometric transformations between images.

A. Proposed architecture

Our proposed visual substitution system is based on the identification of objects around the blind person. We propose a system that recognize and locate 2D in the video (Fig. 1). This system should find the invariant characteristic of objects to viewpoint changes, provide the recognition and reduce the complexity of detection.

We propose a method based on keypoints extraction and matching in video. A comparison between query frame and

database objects is made to detect object in each frame. For each object detected an audio file containing the information about it is activate. Hence object detection and identification are simultaneously addressed.

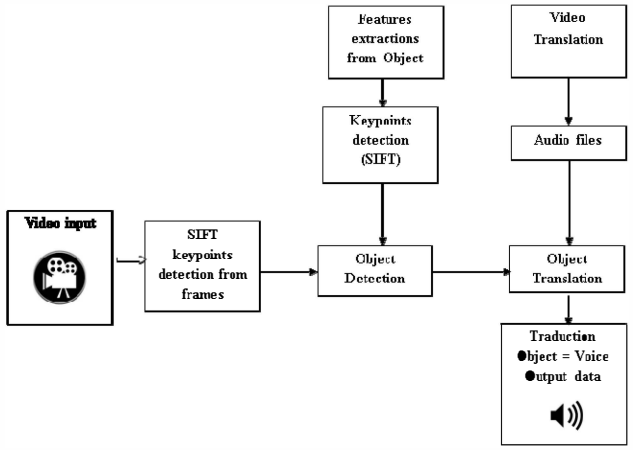

Fig. 1. Flowchart of proposed system

B. Algorithm

We build the bag of key-points of objects with SIFT and we calculate the local dissimilarity map between frames. The algorithm demonstrating the RVLDM used in the video analysis processes as follow:

(1) Obtain the set of key-points of objects
   a. Select a large set of images of daily objects.
   b. Extract the SIFT feature points of all the images within the set and obtain the SIFT descriptor for each feature point extracted from each image.

(2) Obtain the keypoints descriptor for the first video frame.
   a. Extract SIFT feature points of the given image.
   b. Acquire SIFT descriptor for each feature point.
   c. Match the frame key-points with those of the objects and identify detected objects.

(3) For the next frame:
   a. If it contains the same objects they will not be detected
   b. New objects will be detected and identified.
   c. Another method will be used in this step to identify similar and dissimilar frames for further treatments.

(4) For each object detected in video a video file is launched to notify the blind about the identity of objects.

This algorithm highlights the use of SIFT for keypoints detection and matching to ensure the identification of the objects.

V. RESULTS

A. Testing data

Available video databases consist generally on videos of traffic vehicles, faces detection and tracking. In our work we need videos of scenes from daily life with images of all the objects. We precede with four videos sequences (Fig. 2); each one from them contains a number of daily life objects.
Videos have not the same duration. Table I shows that each one contains different number of frames and objects. They will have different time processing.

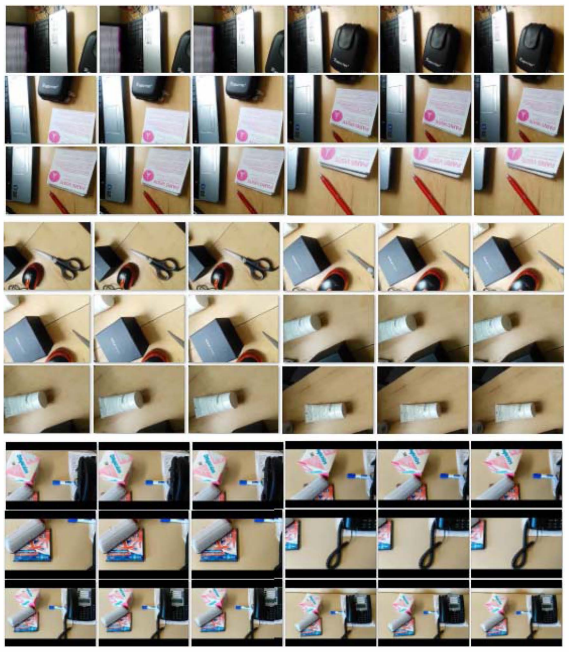

Fig. 2. Example of selected frames from video scenes

TABLE I. DATA SELECTED FOR TESTS

Video	Number of frames	Number of objects
1	1410	6
2	1022	8
3	786	6
4	685	4

B. Object detection and identification in video

The first step is to load up a video scene containing objects; then we have the feature extraction. We use the difference-of-Gaussian feature detector introduced previously with SIFT descriptor. The features we find are described in a way which makes them invariant to size changes, rotation and position. These are quite powerful features and are used in a variety of tasks. We use a standard Java implementation of SIFT [15].
The SIFT descriptor is a 128 dimensional description of a patch of pixels around the keypoint.
In the step of matching, the basic matcher finds many matches, many of which are clearly incorrect.

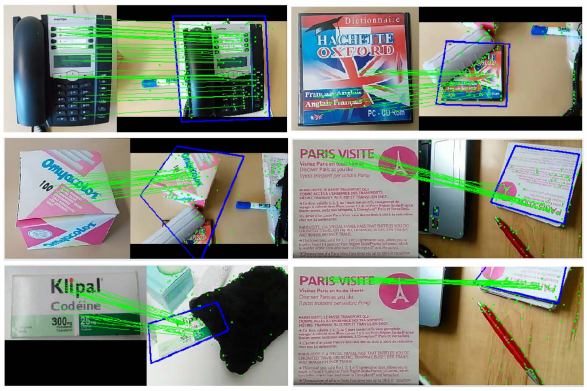

Fig. 3. Object detection in some video scenes

Fig. 4. Detection of a medical box

TABLE II. NUMBER OR TRUE MATCHES AND FALSE MATCHES DEPENDING ON THE SCALE NUMBER

Number of scales	True positive (%)	False positive (%)
3	35	65
5	95	5

An algorithm called Random Sample Consensus (RANSAC) [16] is used to fit a geometric model called an affine transform to the initial set of matches. This is achieved by iteratively selecting a random set of matches, learning a model from this random set and then testing the remaining matches against the learnt model.

We can take advantage of this by transforming the bounding box of the object with the transform estimated in the affine transform model. Therefore we can draw a polygon around the estimated location of the object within the frame.

Then, Table II shows that the number of true positive depends on the number of scales used in the SIFT algorithm. In fact, 5 scales gives a better matching than 3, also a number up to 5 didn't' give a butter results but it takes more time. So, in order to have strong matches, we work with number of scale S equal to 5. In the figure above (Fig.3) we tried to find some object. We note that we detect all objects in this scene video. So we tried also to detect a medical box, because we know that identifying medicines is a delicate task for blind people.

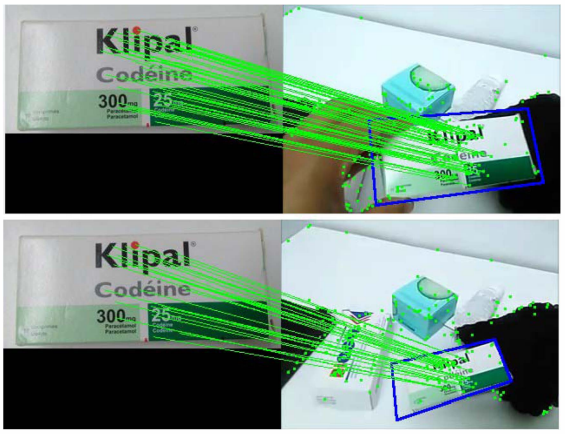

Fig. 5. Medical box detection in video with low luminosity

Fig. 6. Medical box detection in video with high luminosity

In the figure (Fig.4), the box is well detect, a short description about the medicine in an audio file notify the blind about what he hold in his hand.

In the video scene we can have different level of illumination. Even in the same video, illumination can be changed. It was important to detect objects in video scene with high illumination (Fig.5) and in another one with low illumination (Fig.6). It is clear that the number of true matches decrease in the video with low lightness. However, the object is well detected. So we can conclude that SIFT is invariant to the change in luminosity in video and the object can be detected and identified.

C. *Discussions*

The challenge in comparing keypoints is to figure out matching between keypoints from some frames and those from target objects. We get high percentage of detected objects but we tried also to identify the reason behind some failure cases.

The first cause of non-detection of object was the quality of images. For example in the figure (Fig. 7), the pen was not detected because we have added some noise to the video scene.

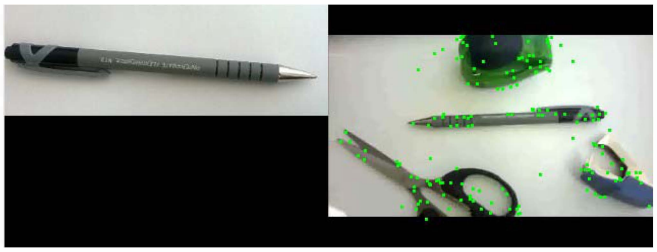

Fig. 7. Failure detection caused by the quality of image

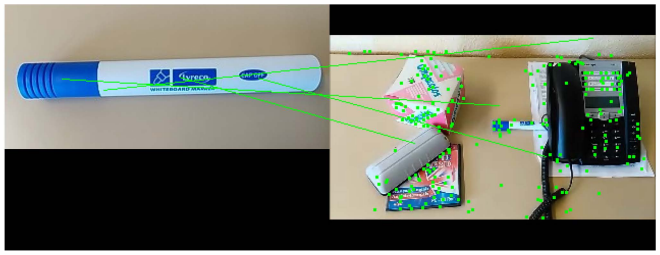

Fig. 8. Failure detection caused by the size of target object

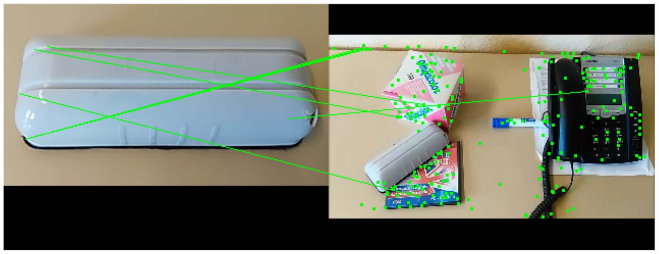

Fig. 9. Failure detection caused by the high speed of video scene

Then in the figure (Fig. 8), the object in the target frame is very small compared to his origin size. We know that SIFT is invariant to the scale transformation but when the object contain some texture and it is very small in the target frame, the number of keypoints is not sufficient for matching object and frame and we get all the time false matches.

Furthermore, when the registered video is very fast, SIFT has not the time to detect all keypoints and we get generally a false matches (Fig. 9) so the object is not well detected.
In some works they have to slow down the video, but in our case, we have just to move slowly in the step of recording the scene.

VI. CONCLUSION AND FUTURE WORKS

In this paper, we present a visual substitution system for blind people based on object recognition in video scene. This system uses SIFTS keypoints extraction and features matching for object identification. We devote the experimental part to test the application in order to detect some objects in some video scene with different conditions.

In this stage of works, we address the recognition of each object in the scene as an individual task; we do not consider the relationships between many objects. Thus, in future works, we will consider this relationship for scene understanding or detecting everything that belongs to a given place or location.
Finally, in order to help bind people and to provide from the new technologies, a mobile application can be the best solution.

REFERENCES

[1] Durette, B., Louveton, N., Alleysson, D., and H´erault, J, 2008. Visuo-auditory sensory substitution for mobility assistance: testing The VIBE. *In Workshop on Computer Vision Applications for the Visually Impaired, Marseille, France.*

[2] Hern´andez, A. F. R. et al, 2009. Computer Solutions on Sensory Substitution for Sensory Disabled People. In Proceedings of the 8th WSEAS International Conference on Computational Intelligence, Man-machine Systems and Cybernetics, pp. 134–138.

[3] Tang, H., and Beebe, D. J., 2006. An oral tactile interface for blind navigation. *IEEE Trans Neural Syst Rehabil Eng, pp. 116–123.*

[4] Auvray, M., Hanneton, S., and O'Regan, J. K., 2007. Learning to perceive with a visuo - auditory substitution system: Localisation and object recognition with 'The vOICe'. *Perception, pp. 416–430.*

[5] Kammoun, S. et al, 2012. Navigation and space perception assistance for the visually impaired: The NAVIG project. *In IRBM, Numro spcial ANR TECSAN, 33(2), pp.182–189.*

[6] Brian Katz, F.G. et al, 2012. NAVIG: Guidance system for the visually impaired using virtual augmented reality. *In Technology and Disability, pp. 163–178.*

[7] Abboud, S., Hanassy, S., Levy-Tzedek, S., Maidenbaum, S., and Amedi, A., 2014. EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution. *In Restorative neurology and neuroscience.*

[8] Martnez, B.D.C., Vergara-Villegas, O.O., Snchez, V.G.C., De Jess Ochoa Domnguez, H., and Maynez, L.O., 2011. Visual Perception Substitution by the Auditory Sense. *In Beniamino Murgante, Osvaldo Gervasi, Andrs Iglesias, David Taniar, and Bernady O. Apduhan, editors, ICCSA (2), volume 6783 of Lecture Notes in Computer Science, pp. 522–533. Springer.*

[9] Jabnoun, H., Benzarti, F., and Amiri, H., 2014. Visual substitution system for blind people based on SIFT description. *In International Conference of Soft Computing and Pattern Recognition. Tunsia, pp. 300–305.*

[10] Renier, L. et al, 2005. The Ponzo illusion with auditory substitution of vision in sighted and early-blind subjects. *In Perception.*

[11] Zhang, J., Ong, S. K. and Nee, A. Y. C., 2009. Design and Development of a Navigation Assistance System for Visually Impaired Individuals. *In Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology, i-CREATe '09, New York, NY, USA.*

[12] Jafri, R., Ali, SA., Arabnia, HR., and Fatima, S., 2013. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *In The Visual Computer, Springer Berlin Heidelberg, pp. 1–26.*

[13] Lowe, G.D., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *In Int. J. Comput. Vision, 60(2), pp. 91–110.*

[14] Bay, H. , Ess, A., Tuytelaars,T., and Gool, L.V. 2008. SURF: Speeded Up Robust Features. Computer Vision and Image Understanding (CVIU), pp. 346–359.

[15] S. Hare., J., Sina,S., and P. Dupplaw, D. 2011. OpenIMAJ and ImageTerrier: Java libraries and tools for scalable multimedia analysis and indexing of images. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 691-694. DOI=10.1145/2072298.2072421.

[16] Ondřej, C.. and Matas, J. 2005. Matching with PROSAC-progressive sample consensus. Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. Vol. 1. IEEE.